LEARNING DISCRIMINATIVE SPARSE MODELS FOR SOURCE SEPARATION AND MAPPING OF HYPERSPECTRAL IMAGERY

By

Alexey Castrodad, Zhengming Xing

John Greer, Edward Bosch

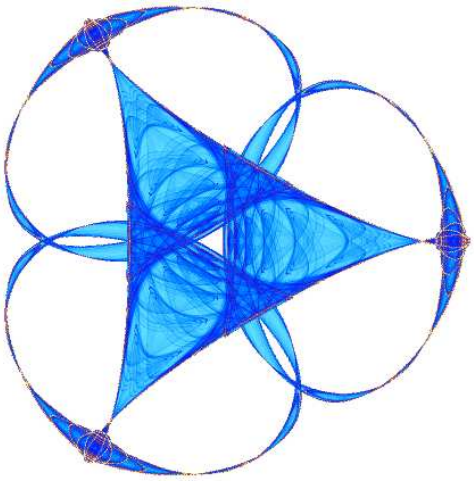
Lawrence Carin

and

Guillermo Sapiro

IMA Preprint Series # 2341

(October 2010)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA

400 Lind Hall
207 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612-624-6066 Fax: 612-626-7370
URL: http://www.ima.umn.edu

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
OCT 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Learning Discriminative Sparse Models for Source Separation and Mapping of Hyperspectral Imagery	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Minnesota,Institute for Mathematics and Its Application,207 Church Street SE,Minneapolis,MN,55455-0436**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A method is presented for sub-pixel mapping and classification in hyperspectral imagery, using learned blockstructured discriminative dictionaries, where each block is adapted and optimized to represent a material in a compact and sparse manner. The spectral pixels are modeled by linear combinations of subspaces defined by the learned dictionary atoms, allowing for linear mixture analysis. This model provides flexibility in the sources representation and selection, thus accounting for spectral variability, small-magnitude errors, and noise. A spatial-spectral coherence regularizer in the optimization allows for pixels classification to be influenced by similar neighbors. We extend the proposed approach for cases for which there is no knowledge of the materials in the scene, unsupervised classification and provide experiments and comparisons with simulated and real data. We also present results when the data have been significantly under-sampled and then reconstructed, still retaining high-performance classification, showing the potential role of compressive sensing and sparse modeling techniques in efficient acquisition/transmission missions for hyperspectral imagery.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**31**	

Learning Discriminative Sparse Models for Source Separation and Mapping of Hyperspectral Imagery

Alexey Castrodad, Zhengming Xing, John Greer, Edward Bosch, Lawrence Carin, and Guillermo Sapiro

Abstract

A method is presented for sub-pixel mapping and classification in hyperspectral imagery, using learned block-structured discriminative dictionaries, where each block is adapted and optimized to represent a material in a compact and sparse manner. The spectral pixels are modeled by linear combinations of subspaces defined by the learned dictionary atoms, allowing for linear mixture analysis. This model provides flexibility in the sources representation and selection, thus accounting for spectral variability, small-magnitude errors, and noise. A spatial-spectral coherence regularizer in the optimization allows for pixels classification to be influenced by similar neighbors. We extend the proposed approach for cases for which there is no knowledge of the materials in the scene, unsupervised classification, and provide experiments and comparisons with simulated and real data. We also present results when the data have been significantly under-sampled and then reconstructed, still retaining high-performance classification, showing the potential role of compressive sensing and sparse modeling techniques in efficient acquisition/transmission missions for hyperspectral imagery.

Alexey Castrodad and Guillermo Sapiro are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, 55455 USA e-mail: {castr103, guille}@umn.edu. Alexey Castrodad, John Greer, and Edward Bosch are with the Department of Defense. Zhengming Xing and Lawrence Carin are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708 USA e-mail:{zhengming.xing, lcarin}@duke.edu.

Learning Discriminative Sparse Models for Source Separation and Mapping of Hyperspectral Imagery

I. Introduction

Hyperspectral imaging (HSI) systems acquire images in which each pixel contains narrowly spaced measurements of the electromagnetic spectrum,[1] allowing spectroscopic analysis. The data acquired by these spectrometers play significant roles in biomedical, environmental, land-survey, and defense applications. It contains the geometrical (spatial) information from standard electro-optical systems, and also much higher spectral resolution, essential for material identification.

There are numerous intrinsic challenges associated with effective ground mapping and characterization applications when using overhead HSI, see for example [1]–[3]. The first is the noise of the collected measurements, which directly affects detection accuracy and sensitivity to low presence of materials. Secondly, the complicated schemes of energy interaction between the targeted area and the spectrometer, causing the total count of photons at the sensor's photoelectric array to include energy from contributing factors from the atmosphere such as aerosols and water vapor. The third challenge occurs at the surface level, where spatial resolution and reflected light off nonuniform surfaces generate intrinsic spectral variability associated with each material, and *spectral mixing*, where each pixel is often composed of a combination of materials. In this case it is difficult to match the data with controlled (laboratory) measured spectra. Finally, the high dimensionality of the data coming from a large number of spectral bands poses challenges in visualization, interpretation, and transmission tasks. Nevertheless, these narrowly spaced bands are highly-correlated and redundant, which allows one to carefully exploit "blessings" of high dimensionality.

In this work we investigate methods that capitalize on this redundancy to address the processing challenges of high-dimensional hyperspectral data. In Section II we provide the necessary tools to get appropriate material representations using sparse coding and dictionary learning techniques. We propose in Section III a material identification scheme at the sub-pixel level, for remotely sensed HSI using a learned block-dictionary optimized to represent user-specific classes in a supervised fashion. After presenting experimental results and comparisons in Section IV, we extend it in Section V to the unsupervised case. The proposed technique efficiently manages data redundancy and provides a representation of the HSI cube as a sparse linear combination of learned sources (dictionary atoms), giving meaningful material abundance estimates. The proposed algorithm does not require explicit dimension reduction or subspace projection preprocessing steps. As we will later see, our methodology has

[1]Typically within the visible and long wave infrared region ($400 - 14000$ nm).

several advantages. First, it gives freedom to each pixel to select an appropriate sparsity level, equivalent to the intrinsic dimensionality and subspace selection, in contrast with standard dimension-reduction algorithms, which assume a global or semi-global dimensionality for all pixels. Second, our proposed classification approach balances between nearest subspace, defined by the learned dictionary atoms, and nearest neighborhood classifiers, where the pixels' abundance coefficients come from a union of subspaces corresponding to all available materials. Third, we impose spatial coherence in the sparse modeling-based classification. This efficiently combines spectral and spatial information by incorporating local and nonlocal connectivity, leading to a grouping criteria that induces a robust and more stable sparse coding (abundance mapping). The proposed algorithm is tested using synthetic and real data, and compared with other mapping techniques in sections IV and VI, where we also show that accurate material identification can be attained even from highly under-sampled data following reconstruction using a probabilistic framework for compressive sensing [4].

Sumarizing, we make the following main contributions:

- We present a new model for supervised and unsupervised spectral source separation with class-specific spectral dictionaries (similar to endmembers) that takes advantage of the data redundancy and accounts for high variability in the materials' spectral response.
- We address the problem of data acquisition, storage, and transmission by showing that accurate material identification can be attained with the proposed approach even under very low sampling conditions.

II. HSI Mapping via Sparse Reconstruction

We seek a technique capable of identifing the set of materials that best represents each pixel on a given hyperspectral image. This is, as mentioned above, a non-trivial problem because of the many factors associated with land remote sensing. We begin with the supervised case, where we know *a priori* (or at least have a good idea of) what materials might be present in the scene, and we use this information to make a composition mapping of the scene. Then, we address the case for which there is no *a priori* information about these materials, the unsupervised case.

Let each measured pixel $\mathbf{y} = [y_1, y_2, ..., y_b]$ in the hyperspectral image be a vector valued function, $y_i : \Re^2 \rightarrow \Re, 1 \leq i \leq b$, where b denotes the number of spectral bands. We stack these pixels in matrix format as $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \Re^{b \times n}$, where n is the total number of available pixels distributed spatially. As previously mentioned, there are several challenges accompanying HSI that jeopardize precise material identification. HSI is noisy in nature, so there are differences between the true and the observed signals. In addition, there are distortions associated with atmosphere suppression models [5] (which convert radiance into reflectance units).[2] We assume that the measured energy $\mathbf{Y}$ at the sensor is proportional to the area covered by the (learned) dictionary of materials $\mathbf{\Psi}$ and the reflectivity of the media, which can be modeled as the linear system

$$\mathbf{Y} = \mathbf{\Psi}\mathbf{A} + \mathbf{N},$$

[2]Our model is not restricted to work with reflectance units.

where $\mathbf{N}$ is additive noise with bounded energy ($\|\mathbf{N}\|_F^2 \leq \sigma^2$), $\mathbf{\Psi} \in \Re^{b\times k}$ is a dictionary (soon to be learned), and $\mathbf{A} \in \Re^{k\times n}$ is the associated matrix of coefficients representing the mixture of dictionary atoms when composing the data. The goal of this work is to learn the dictionary $\mathbf{\Psi}$ representing the materials, and their proper combination $\mathbf{A}$, just from $\mathbf{Y}$ (unsupervised case) or from $\mathbf{Y}$ and a labeled library of real data (supervised case). We will also work with significantly under-sampled data $\mathbf{Y}$.

A. Class-Specific Reconstruction

Suppose there are C known classes, each with an associated label $j \in [1, C]$. These classes are represented by sets of k_j basis vectors of the dictionary, or *dictionary atoms*, such that for the $j-th$ class,

$$\mathbf{\Psi}^j = [\boldsymbol{\psi}_1^j, ..., \boldsymbol{\psi}_{k_j}^j] \in \Re^{b\times k_j}, \forall j \in [1, C].$$

The full dictionary Ψ is composed of the concatenation of these class-dependent dictionaries Ψ^j. The set of n_j pixels $\mathbf{Y}^j \in \Re^{b\times n_j}$ associated with class j has a corresponding matrix of coefficients $\mathbf{A}^j \in \Re^{k_j\times n_j}$ such that ($\boldsymbol{\alpha}_i^j \in \Re^{k_j}$, $i = 1, ..., n_j$)

$$\mathbf{A}^j = [\boldsymbol{\alpha}_1^j, ..., \boldsymbol{\alpha}_{n_j}^j], \forall j \in [1, C].$$

In other words, each column $\boldsymbol{\alpha}_i^j$ in $\mathbf{A}^j$ represents the coding coefficients for one of the pixels in $\mathbf{Y}^j$ (while the upper index stands for the class, the lower index stands for the pixel, running from 1 to n_j for the class or 1 to n for the whole data). The full matrix $\mathbf{A}^j$ then represents all the coefficients needed for the representation of the class pixels $\mathbf{Y}^j$.

Thus, given C classes, we let the $j-th$ class reconstruction cost function for a signal $\mathbf{y}$ from the set of data points $\mathbf{Y}$, be the minimization of (with respect to $\boldsymbol{\alpha}^j$)

$$\mathcal{R}_{\ell_2}(\mathbf{y}, \mathbf{\Psi}^j) = \|\mathbf{y} - \mathbf{\Psi}^j\boldsymbol{\alpha}^j\|_2^2, \text{ s.t. } \boldsymbol{\alpha}^j \succeq 0, \tag{1}$$

where the nonnegativity constraint to $\boldsymbol{\alpha}^j$ is to avoid a "negative influence" from the material represented by $\mathbf{\Psi}^j$ (a$\succeq$b is an elementwise inequality).

For now, assume that the class (material) representative $\mathbf{\Psi}^j$ is known in advance, for all the classes (all j). These class representatives may or may not be part of the image. For example, $\mathbf{\Psi}^j$ may correspond to a class representation derived from the image itself, from another image, or a more generic one such as a spectral library (where for the $j-th$ class, $k_j = 1$). The coefficients in the matrix $\mathbf{A}^j$ quantify the contribution onto each pixel in $\mathbf{Y}$ from each of these classes represented in $\mathbf{\Psi}^j$ (since the coefficients are nonnegative). Assuming $k_j = 1$ (a class represented by a single atom), one would expect that the label associated with the largest coefficient of the minimal reconstruction error is the most dominant material in the pixel. A simple way for classifying each pixel, to its dominant class, after minimizing (1) with respect to $\boldsymbol{\alpha}^j$ is by defining a mapping function $f(\mathbf{y}) : \Re^b \rightarrow [1, C]$ as

$$f(\mathbf{y}) = \{j | \mathcal{R}_{\ell_2}(\mathbf{y}, \mathbf{\Psi}^j) \leq \mathcal{R}_{\ell_2}(\mathbf{y}, \mathbf{\Psi}^i), i \neq j, \forall i, j \in [1, C]\}. \quad (2)$$

Note that (2) can be viewed as a minimum (squared) Euclidean Distance (ED) classifier. The performance of this classifier highly depends on how well the classes/materials are represented in $\mathbf{\Psi}$. Moreover, it may occur that the number of columns in $\mathbf{\Psi}$ is larger than the number of bands, to allow for a rich representation of the class/material, and hence the possibility of many solutions, a matter we will later address by adding more constraints.

Assuming that the class is being represented by a single vector ($k_j = 1$) has several limitations. If Ψ is a spectral library, gross errors could emerge from the interpolation procedure to adapt the spectral library to the data (or vice-versa), or as previously mentioned, from atmospheric correction algorithms. In addition, there are numerous instances of materials encountered in groups. This is typically found in vegetation fields, where there is often a mixture of soil, leaves, and wood, an issue discussed for example in [1], [6], [7]. Therefore, using a single spectra in these cases may not be sufficient to represent the class and its expected variability. One important aspect of high-dimensional spaces is that it is very difficult to completely characterize the classes of interest; i.e., the use of averaged spectral curves or a single spectra (e.g., from a spectral library) is a limiting factor in the performance of the classifier (see [8], [9]). With increasing availability of training samples, the problem should become "easier," since the class should be able to be better studied and represented. In cases where the data cannot be accurately characterized by a single averaged spectra, a better representation should be pursued.

If we let $k_j > 1$ in $\mathbf{\Psi}^j$, we could use all available training data and try to reconstruct the image using this collection of spectra. For a valid reconstruction, we would need the training data to span the subspace where the class lives. In most cases, there is not enough information to do this. This discouraging result could be seen from [10], which states in our case that for an estimator of any Lipschitz function on $[0, 1]$, $\sup E\{(\mathbf{\Psi}\boldsymbol{\alpha} - \mathbf{y})^2\} \geq K \cdot n^{-\frac{2}{(2+b)}}, \quad n \to \infty$, where $E\{\cdot\}$ is the expectation operator, and K is a positive constant. This tells us that the MSE is lower bounded by a very slow convergence rate in the number of samples relative to the dimension (b-channels). Thus, we would think that we need an unrealistic amount of samples per class for a reasonable performance, the so called "curse of dimensionality." However, it is well known that in high-dimensional spaces, the data tends to concentrate in the corners of an hypercube (or shell of a hypersphere), so the volume it spans is mostly empty. This phenomena motivates the search for an optimal subspace where the data lies. In other words, the intrinsic local dimension of the data is expected to be low, giving hope for learning efficient representations from reasonable amounts of data.

There exist very efficient methods for dimensionality reduction and subspace projection, though only a few guarantee that all critical information for class separation is preserved. In other words, finding an optimal subspace at a global level might obviate some details in the data that could be essential for class identification. Linear methods tend to capture an optimal subspace at a global level, and often fail to preserve the local features. Manifold learning and kernel method techniques, including multiple versions of Diffusion Maps, LLE and ISOMAP (see for example [11]–[16]), have been used to improve the capability of feature extraction in HSI. These techniques try to capture local structure information, but often require a high number of operations (typically quadratic in the number

of data samples), and suffer from other difficulties like the "out-of-sample" issue [17]. In addition, it becomes difficult to interpret the resulting transformation in a physical way. We would like to find a representation for each pixel that could efficiently pursue the appropriate subspace/s, while only losing the least essential information. For that matter, we do not seek a global dimension reduction approach. We will exploit the data itself and treat it with model selection techniques that efficiently seek a good trade-off between global and local information.

Furthermore, we will assume that while the amount of different materials in the scene can be large, only a small quantity of these materials is observed in each pixel, and each material itself lives in a low dimensional subspace. Thus, mapping HSI pixels is a sparse coding problem, only a few atoms per class-dictionary $\mathbf{\Psi}^j$ are active at a time (defining the class subspace), and only a few classes j are present per pixel. In [18] the authors showed that if the output of the classification problem is sparse (only a few of many possible classes is present at a time), then with overwhelming probability the amount of data needed to learn is significantly reduced. This gives us further hope into dealing with a relatively low number of training samples while still retaining high classification performance. Minimizing (with respect to $\boldsymbol{\alpha}^j$) the class reconstruction term (1) is not enough to accomplish this, since it will typically induce dense solutions, large support of nonzero coefficients. We take care of this sparsity modeling by the use of a shrinkage-inducing complexity term as detailed next.

B. Accounting for Sparsity in the Reconstruction Coefficients

Sparse modeling has proven over the last decade to be a powerful technique in signal processing. Its goal is to achieve signal representation using the minimum amount of data while retaining the maximum amount of information. Sparse coding (assuming given dictionaries for the moment) comes from the solution to an under-determined systems of equations,

$$\min_{\boldsymbol{\alpha}^j} \|\boldsymbol{\alpha}^j\|_0 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{\Psi}^j \boldsymbol{\alpha}^j, \tag{3}$$

where $\|\cdot\|_0$ is a pseudo-norm that counts the number of nonzero entries. This is a combinatorial problem, usually approximated using greedy methods like Orthogonal Matching Pursuit (MP) [19]. It is well known that (3) is equivalent to solving (under assumptions on the sparsity of the signal and the structure of the dictionary $\mathbf{\Psi}^j$, see [20] and references therein)

$$\min_{\boldsymbol{\alpha}^j} \|\boldsymbol{\alpha}^j\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{\Psi}^j \boldsymbol{\alpha}^j. \tag{4}$$

Since HSI data are noisy, we are interested in a modification of (4), that allows sparse representations under bounded noise, known as basis pursuit denoising [21], or the Lasso [22]. Its nonnegative version is to minimize (with respect to $\boldsymbol{\alpha}^j$)

$$\mathcal{R}_{\ell_{2,1}}(\mathbf{y}, \mathbf{\Psi}^j) = \|\mathbf{y} - \mathbf{\Psi}^j \boldsymbol{\alpha}^j\|_2^2 \quad \text{s.t.} \quad \mathcal{S}(\boldsymbol{\alpha}^j) \leq t, \tag{5}$$

where $\mathcal{S}(x) = \sum_i x(i)$, and $t \in \Re_+$ is a threshold that controls the sparsity of the coefficients (smaller t will cause more shrinkage in the coefficients). Equivalently, we could express (5) in Lagrangian form as

$$\mathcal{R}_{\ell_{2,1}}(\mathbf{y}, \mathbf{\Psi}^j) = \|\mathbf{y} - \mathbf{\Psi}^j \boldsymbol{\alpha}^j\|_2^2 + \lambda_s \mathcal{S}(\boldsymbol{\alpha}^j), \tag{6}$$

where λ_s is a nonnegative penalty term that balances the complexity and the fitting quality (recall that the coefficients in α^j are constrained to be non-negative). Note that each set of selected atoms (non-zero entries of $\boldsymbol{\alpha}^j$) from $\mathbf{\Psi}^j$ defines a subspace, and therefore the class/material is represented by a collection of such subspaces, which are low dimensional due to the sparsity constraint.

Having performed the above optimization for each class (with standard optimization techniques), we can then classify our data according to (6), assigning to each pixel the class label corresponding to minimum energy (we assume for the moment a single class per pixel), thus the mapping

$$f(\mathbf{y}) = \{j | \mathcal{R}_{\ell_{2,1}}(\mathbf{y}, \boldsymbol{\alpha}^j) \leq \mathcal{R}_{\ell_{2,1}}(\mathbf{y}, \boldsymbol{\alpha}^i), i \neq j, \forall i, j \in [1, C]\}. \tag{7}$$

Here the model complexity also serves as a discriminative part of the classifier. Specifically, the pixel will be assigned to the class with minimum reconstruction error and minimum complexity in the coefficient vector. This classifier has been proposed and used in [23] for texture and handwriting classification, showing advantages of using both terms instead of only using the reconstruction term as in previous approaches.

C. Dictionary Learning

Originally sparse representations were obtained using fixed "off-the-shelf" dictionaries $\mathbf{\Psi}^j \in \Re^{b \times k_j}$, where k_j is the number of atoms, and b is the signal's dimension (e.g., DCT, Fourier basis, wavelets). It is often more appropriate to "learn" these dictionaries by adapting them to the data at hand [24]–[28]. State-of-the-art results have been reported in applications related to noise removal, inpainting, discriminative learning from image databases, classification, and unsupervised labeling (clustering) [23], [29]–[34].

The dictionary is learned by minimizing (with respect to both $\mathbf{\Psi}^j$ and $\mathbf{A}^j$)

$$\mathcal{R}_{\ell_{2,1}}(\mathbf{Y}^j, \mathbf{A}^j) = \|\mathbf{Y}^j - \mathbf{\Psi}^j \mathbf{A}^j\|_F^2 + \lambda_S \sum_i^{n_j} \mathcal{S}(\boldsymbol{\alpha}_i^j). \tag{8}$$

This process of simultaneously learning a dictionary and obtaining a sparse representation is called *sparse modeling* (in contrast to *sparse coding*). This is a non-convex optimization problem usually solved by an alternated minimization scheme, that is, fixing $\mathbf{A}^j$ while updating $\mathbf{\Psi}^j$ and vice-versa. The main idea is to break this minimization in two parts. First, fix the dictionary atoms and perform sparse coding (e.g., using an algorithm like Least Angle Regression and Shrinkage [35]). Then one fixes these sparse coefficients and updates the atoms of the dictionary. There are several ways for learning these atoms. The K-SVD and Method of Orthogonal Directions (MOD) algorithms [25], [36] are widely used in the processing of natural images for this task. In this work, we use a Projected Gradient (PG) iteration, where we update the $i-th$ atom $\boldsymbol{\psi}_i^j$ by the following scheme:

$$\begin{aligned}\boldsymbol{\psi}_i^{j,t} &\leftarrow P\{\boldsymbol{\psi}_i^{j,t-1}\} \\ &= \max(0, \boldsymbol{\psi}_i^{j,t-1} + \mu(\frac{\partial}{\partial \boldsymbol{\psi}_i^j}\mathcal{R}_{\ell_2})) \\ &= \max(0, \boldsymbol{\psi}_i^{j,t-1} + \mu(2(\boldsymbol{\Psi}^j\mathbf{A}^j - \mathbf{Y})\mathbf{A}_i^{jT}),\end{aligned}$$

where $\mathbf{A}_i^j$ denotes the $i-th$ row of $\mathbf{A}^j$, and $\mu = \frac{0.9}{\|\mathbf{A}^j\mathbf{A}^{jT}\|}$ is the step-size parameter, where $\|\cdot\|$ is the spectral norm. Each dictionary atom is normalized to have unit norm, projected onto the unit sphere after each iteration.

There are a number of possible advantages of using a compact learned dictionary instead of all the the data samples as the dictionary. The first advantage is seen from (8), where $\boldsymbol{\Psi}^j$ is modified such that the actual reconstruction error is minimized. Secondly, a learned dictionary provides a more compact way of representing a class, $k_j \ll n_j$, reducing the sparse coding computational cost (note that in the supervised learning case for example, the dictionary is learned once off-line).

III. Multi-label Prediction for Abundance Mapping

Low spatial resolution distorts the geometric features in the scene, and introduces the possibility of having multiple materials inside a pixel. In addition, partial occlusions caused by elevation differences will also cause such mixtures. For example, if there are tree branches over a road in the scene, the measured pixels are a combination of the energy reflected from the tree leaves and from the partially occluded road. Therefore, in general, the pixels in the acquired scene are *not* pure. This effect is known as *spectral mixing*. In general, one may assume that the pixels in the scene contain mixtures of multiple materials. Therefore, a more realistic approach to HSI classification is to allow for a pixel to have one or more labels, each corresponding to a material class. We now extend the above classification scheme to address this more general scenario.

The spectral mixing problem can be seen as a particular case of source separation, where signatures from pure materials are combined to produce the pixel's spectral response. Spectral mixing is caused by a variety of physical factors, many of them often occurring in a nonlinear fashion, and are often difficult to physically model or require many *in situ* parameters associated with the characteristics of the target and the environmental conditions at the time of acquisition. Thus for simplicity of development, and as commonly done in the literature [37], we focus on a linear mixing model (LMM), that is, each pixel is represented as a linear combination of sources or *endmembers*. The coefficients associated with each of these endmembers are called the *fractional abundances*. These fractional abundances indicate the contribution of the associated endmember to the analyzed pixel.[3]

One of the most used models for HSI source separation is the Constrained Least Squares Model (CLS), where the nonnegative coefficients associated with each pixel are constrained to sum to one; see [37] and references therein for more details on the CLS and other proposed models. It is also desirable that the abundance vectors be

[3]Since the learned dictionaries play the role of endmembers in classical HSI analysis approaches, we often call the dictionary and its atoms "endmemebers" as well, even though these are not necessarily pure materials.

sparse, meaning that the material at each pixel is explained with as few as possible pure sources. The sum-to-one constraint of the non-negative coefficients (i.e., that the coefficient vector lines on a simplex), in the CLS model is known to induce a sparse solution, hence a fixed ℓ_1 norm on every coefficient vector. More recently, the Least Squares ℓ_1 (LSℓ_1) model was proposed for this spectral unmixing problem [38], [39]. In this model, the sum-to-one constraint was relaxed, meaning an ℓ_1-norm constraint on the abundance coefficients needs to be minimized, instead of summing strictly one (no learned dictionary is exploited in these works).

An extension to the problem of spectral mixing can be naturally formulated using the concepts from Section II, adapting them to the LMM. Here $\mathbf{\Psi}^j$ represents the $j-th$ material and $\mathbf{A}^j$ its corresponding abundances. Compared to the standard LMM, where the endmembers are real spectral signatures, here the endmembers are represented as a set of subspaces, thus are learned atoms and their combinations and not actual pure materials. This gives the flexibility to account for material variability caused by factors like noise and non-homogeneous substances. In addition, the pixels pertaining to a certain class have the flexibility to (sparsely) select the corresponding material atoms that best fit them, and thus more degrees of freedom for a better reconstruction/representation, still with a compact representation. Representing endmembers with more than one representative vector has been used for example in [6], suggesting that endmembers (especially in vegetation) should be represented by a set of spectra, where the abundances were calculated for each element of this set.

The main idea is to train a dictionary for each class, and then form a block-dictionary $\mathbf{\Psi} = [\mathbf{\Psi}^1, ..., \mathbf{\Psi}^C] \in \Re^{b \times k}$, where $k = \sum_{j=1}^{C} k_j$. In this way, the sparse coding on each pixel comes from a sparse "mixed" union of subspaces. In this work, we use the constrained sparse coding step of (6), and select a λ_S such that it gives an ℓ_1 norm of the non-negative coefficient vectors close to one (typically in the $[0.9, 1.1]$ range). This relaxation avoids the need to introduce a zero vector included as an endmember (zero-shade endmember), and therefore allowing shade and dark pixels to be accounted for [40], while also giving sparser results.

After the sub-dictionaries $\mathbf{\Psi}^j$ forming $\mathbf{\Psi}$ have been learned, one per class j, the proposed method for HSI classification and abundance mapping solves the following optimization problem:

$$\min_{(\alpha_i \in \Re^k) \succeq 0} \sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{\Psi}\boldsymbol{\alpha}_i\|_2^2 + \lambda_S \sum_{i=1}^{n} \mathcal{S}(\boldsymbol{\alpha}_i) + \lambda_G \sum_{i=1}^{n} \mathcal{G}(\boldsymbol{\alpha}_i, \mathbf{w}_i), \tag{9}$$

where the first two terms account for reconstruction and sparsity, respectively, and the third term accounts for a grouping and coherence constraint, which is explained next. With this framework, the ℓ_1 energy of the $\boldsymbol{\alpha}_i$ coefficients corresponding to the block $\mathbf{\Psi}^j$ in $\mathbf{\Psi}$ indicates the amount of material from the $j-th$ class in the mixture for the pixel i. Similarly, we can use the reconstruction limited to atoms and coefficients of a given class to determine the contribution of that class to the pixel i.

A. Imposing Spatial Coherence

Up to this point, each pixel was treated independently from each other ($\lambda_G = 0$ in the equation above). To exploit the geometric structure in the image, one can make the estimation of the abundance coefficients $\boldsymbol{\alpha}_i$ for a given

pixel to be influenced by neighboring pixels, introducing spatial and spectral coherence in the modeling and coding process. This coherence will depend both on the pixels' spectral shape *and* the coefficient vector similarities. This can be implemented by defining a function $\mathcal{G}$ that behaves as a grouping (coupling) term on the coefficients,

$$\mathcal{G}(\boldsymbol{\alpha}_i, \mathbf{w}_i) = \|\mathbf{M}(\boldsymbol{\alpha}_i - \sum_{l\in\eta} w_{il}\boldsymbol{\alpha}_l)\|_2^2, \tag{10}$$

where η denotes the neighborhood of the $i-th$ pixel. We define a weighting function $w_{il} = \frac{1}{C_i}\exp\frac{-\|\mathbf{y}_i-\mathbf{y}_l\|_2^2}{\sigma^2}$, where C_i is a pixel-dependent normalization constant, such that $\sum_{l\in\eta} w_{il} = 1$, and σ^2 is a density parameter controlling the width of the Gaussian (here set to be the average of the data pairwise Euclidean distance, either local for each pixel or global for the whole data). This weighting function is close to 1 if the pixels are very similar and 0 if orthogonal. Its purpose is to compare the $i-th$ pixel with a weighted linear combination of its neighbors. There is no guarantee that pixels with strong similarities will select the same active set (atoms) from the $j-th$ class sub-dictionary $\boldsymbol{\Psi}^j$, even in cases where they have similar mixtures. We do want these coefficient vectors to be coupled by similar ℓ_1-norm in each block. For this purpose, $\mathbf{M} \in \Re^{C\times k}$ is defined as $\mathbf{M} = \begin{bmatrix} \mathbf{1}^1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}^2 & & \\ \vdots & & \ddots & \\ \mathbf{0} & \cdots & & \mathbf{1}^C \end{bmatrix}$, where $\mathbf{1}^j \in \Re^{1\times k_j}$ is a vector of ones corresponding to the number k_j of atoms per sub-dictionary. The purpose of $\mathbf{M}$ is to compare the similarity, in the ℓ_1 norm, of the coefficients from each sub-dictionary, promoting similar per-block ℓ_1 norm for similar pixels.

The neighborhood η is not restricted to spatial neighbors, but also non-local neighbors with strong similarities. Let the weights be represented in the matrix $\mathbf{W} \in \Re^{n\times n}$. This weight matrix is formed by the sum of local and nonlocal weights, $\mathbf{W} = \mathbf{W}_{nl} + \mathbf{W}_s$, where $\mathbf{W}_{nl}$ is a weight matrix associated with similar abundance vectors, and $\mathbf{W}_s$ is the weight matrix associated with the spatial neighbors; $\mathbf{W}$ is similar to the matrix used in the nonlocal means algorithm [41] and the spectral/spatial kernel matrix approach of [42]. Incorporating this grouping term gives robustness to noise and stability in the coefficient (abundances) estimates, capturing non-linearities and complicated structures in the feature space. Note that these weights can be calculated prior to optimization (standard techniques to accelerate non-local means can be used if desired).

The optimization problem in (9) can no longer be solved independently for each pixel due to the coupling in the coefficient vectors. However, we efficiently solve this by considering ($\mathbf{I}$ is an identity matrix)

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \mathbf{Y} \\ \lambda_G \mathbf{W}\mathbf{A} \end{bmatrix}, \quad \tilde{\boldsymbol{\Psi}} = \begin{bmatrix} \boldsymbol{\Psi} \\ \lambda_G \mathbf{I} \end{bmatrix}.$$

We solve the coupling using a standard Gauss-Seidel type of iteration (or primal decomposition), where we iteratively solve the problem by first calculating the sparse coefficients with no coupling, storing a copy, and then re-calculating the subsequent coding iteration using this copy.

This concludes the model for supervised HSI classification and unmixing. We next present experimental results supporting this proposed framework.

IV. Supervised Source Separation Experiments

We consider a series of experiments to test the performance of the proposed algorithm. We use three HSI scenes described below. We compare our proposed framework with standard methods for HSI classification, that is, an Euclidean Distance classifier (ED) and a Spectral Angle Mapper classifier (SAM), see below for the exact definitions. In addition, we include an ℓ_1-based optimization scheme that uses all available training samples as dictionary (see next for a detailed description of where the training samples come from), hence a dictionary of data samples of much larger cardinality than our compact learned dictionary. Our algorithm is termed the Dictionary Modeling (DM) when no spatial coherence is imposed, and Dictionary Modeling with Spatial Coherence (DMS) when the spatial coherence is imposed. For all the following experiments, we set the algorithm parameters to be $\lambda_S = 0.5/\sqrt{b}$, and $\lambda_G = 0.01$ (only for DMS). The weight matrix $\mathbf{W}$ was built using a 3×3 spatial region surrounding the pixel, and 4 nonlocal neighbors. A maximum number of iterations for the dictionary learning phase was set to 150, and 5000 for the sparse coding phase. The results were found to be stable to the particular selection of these parameters, which can in general be done via standard cross-validation. We selected the sparse coding technique called Least Angle Regression and Shrinkage (LARS) [35], and used the fast implementation in the Sparse Modeling Software (SPAMS), publicly available at http://www.di.ens.fr/willow/SPAMS/. For DM, DMS, and ℓ_1 methods, all pixels were normalized to have unit norm prior to processing.

A. HSI Data Sets

We process 3 hyperspectral scenes for these experiments, Figure 1:

1) AVIRIS Indian Pines: The Indian Pines is a small portion of the Northwest Tippecanoe County in Indiana, acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) system in 1992. It consists of a $145 \times 145 \times 220$ datacube reduced to 188 bands after removing water absorption and noisy bands. The data are publicly available at https://engineering.purdue.edu/ biehl/MultiSpec/hyperspectral.html. The scene consists mainly of 16 classes, mostly vegetation and agricultural crops, and a full scene classification ground-truth is available.
2) HyDICE Urban: The Urban scene was acquired with the Hyperspectral Digital Collection Experiment (HyDICE) sensor over Copperas Cove, Texas. It consist of a $307 \times 307 \times 210$ datacube reduced to 162 channels after removing the water absorption and noisy bands. 8 classes were manually selected. The data are publicly available at http://www.agc.army.mil/hypercube/.
3) HyMAP APHill: The APHill scene was taken with the Hyperspectral Mapper (HyMAP) over Virginia (with permission from the US Army Engineer Research and Development Center, Topographic Engineering Center, Fort Belvoir, VA). It consists of a $645 \times 400 \times 128$ datacube, reduced to 106 channels after removing noisy and water absorption bands. Nine classes were manually selected for the experiments.

Table I summarizes the class and training/testing samples information for each of these datacubes.

(a) (b) (c)

Fig. 1: HSI cubes used in this work. (a) False RGB of AVIRIS Indian Pines: no patches, all the scene has ground-truth. (b) HyDICE Urban: patch color corresponds to each of the 8 known classes. (c) HyMAP APHill: patch color corresponds to each of the 9 known classes. This is a color figure.

B. Experiment 1: Supervised Multi-label Mapping

In the first experiment, we look at the overall training accuracy for the three datasets. We compare our proposed algorithm with two classical approaches, minimum Euclidean Distance (ED) and Spectral Angle Mapper (SAM). We also compare this approach with an ℓ_1 minimization scheme without learned dictionaries. We define these classifiers as

1) Euclidean Distance:
   $f_{\text{ED}}(\mathbf{y}) = \{j | \|\mathbf{y} - \hat{\mathbf{y}}^j\|_2 \leq \|\mathbf{y} - \hat{\mathbf{y}}^i\|_2, i \neq j, \forall i, j \in [1, C]\}$, where $\hat{\mathbf{y}^j}$ is the averaged spectra of all training samples from the $j-th$ class.
2) Spectral Angle Mapper: The SAM is defined as $\text{SAM}(\mathbf{x}, \mathbf{y}) = \cos^{-1}(\frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2})$, hence the mapping is $f_{\text{SAM}}(\mathbf{y}) = \{j | \text{SAM}(\mathbf{y}, \hat{\mathbf{y}}^j) \leq \text{SAM}(\mathbf{y}, \hat{\mathbf{y}}^i), i \neq j, \forall i, j \in [1, C]\}$.
3) ℓ_1 minimization: Here we use an approximate sparse solution to (4), $\boldsymbol{\alpha}^* = \text{argmin}_{\boldsymbol{\alpha} \geq 0} \|\boldsymbol{\alpha}\|_1$ s.t. $\|\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\alpha}\|_2^2 \leq \eta$, for some small error $\eta = 0.05$. Here $\boldsymbol{\Psi}$ is a matrix with all the training samples available from all classes, each block $\boldsymbol{\Psi}^j$ corresponding to the samples from class j (concatenated to create $\boldsymbol{\Psi}$). We assign a label following the mapping $f_{\ell_1}(\mathbf{y}) = \{j | \mathcal{R}(\boldsymbol{\alpha}^j) \leq \mathcal{R}(\boldsymbol{\alpha}^i), i \neq j, \forall i, j \in [1, C]\}$, where $\boldsymbol{\alpha}^j$ is the coding portion corresponding to the sub-dictionary $\boldsymbol{\Psi}^j$ for class j and $\mathcal{R}$ is the reconstruction error when coding the data sample (pixel) only with samples selected by $\boldsymbol{\alpha}^j$ (equivalent to reconstruction after setting to zero all elements of $\boldsymbol{\alpha}^*$ but those in $\boldsymbol{\alpha}^j$). Similar results are obtained with $\mathcal{S}$ (the ℓ_1 norm) instead of $\mathcal{R}$.

We divided the set of known samples from each class into training and testing sets by random selection for 25 draws. We selected the number of atoms per sub-dictionary to be $k_j = \min(50, n_j)$, where n_j is the total number of training samples for the $j-th$ class. The results are summarized in Table II, showing the average and standard deviation of the overall classification accuracy for the 25 runs. We make the following observations:

1) The results show that using the compact learned dictionary in the proposed framework outperforms the other more classical approaches. Compared to using the data itself as dictionary, a significantly reduced computational cost is obtained as well.
2) The spatial coherence term significantly improved the classification accuracy in the Indian Pines dataset, mainly due to the large uniform areas. Also, the ED and SAM methods performed poorly on this data, using averaged spectra was not sufficient for a good class representation. In contrast, the proposed approach, while simple in nature, efficiently selects the atoms that best represent the rich classes.
3) Although the classes seem to be more "separable" in the APHill and Urban datasets, leading to relatively good results using ED and SAM, the proposed approach provides close to 100% accuracy, which could play an essential role in certain mapping applications.

C. *Experiment* 2*: Mapping of Reconstructed Data From Significant Under-sampling*

Recently, a non-parametric (Bayesian) approach to sparse modeling and compressed sensing was proposed in [4], where most of the data is eliminated uniformly at random, and then reconstructed using a dictionary that is learned only from the available data. The method automatically estimates the dictionary size, and makes no explicit assumption on the noise variance. In addition, it can deal with non-uniform noise sources in the different bands, a problem often encountered in HSI. While the method can reconstruct the data with a high Peak Signal to Noise Ratio (PSNR), we investigate how well the spectral information is preserved, that is, we test how material classification is affected after randomly "throwing away" most of the data and then interpolating. In Table III we provide an idea on how this subsampling and reconstruction affects the spectral angles between the original and reconstructed pixels. Most of the high angles are due to two reasons: first, small spatial areas could not be appropriately reconstructed, since the approach uses information from neighboring pixels (spatial window) for interpolation. Secondly, areas with low SNR fail to be efficiently reconstructed. As an example, the lakes in the APHill scene have the highest spectral angles, mainly because most of the energy from the sun was absorbed by the water. These problems can be addressed by an adaptive sampling strategy, instead of the random one here employed for testing the classification accuracy.

Table IV summarizes the classification accuracies obtained for the HSI datasets under several under-sampling conditions. We compare our proposed algorithm with the same ones used in Experiment 1. We make the following observations:

1) The proposed DM/DMS algorithms outperform the other methods in most cases. However, there are some cases with the IndianPines images were the ℓ_1 minimization method with data as dictionary performs slightly better, particularly in extreme cases were only 2% of the data is used. This difference in performance was mainly on classes with a small number of training samples (e.g., Alfalfa, Oats, and Grass/pasture-mowed classes). The learned dictionary was not able to appropriately model the class (this can be easily solved letting the dictionary be the data itself when not enough training samples are available). Nevertheless, this problem was solved with

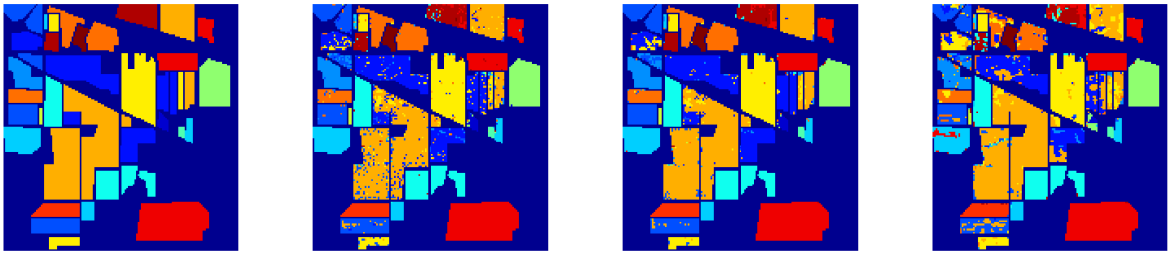
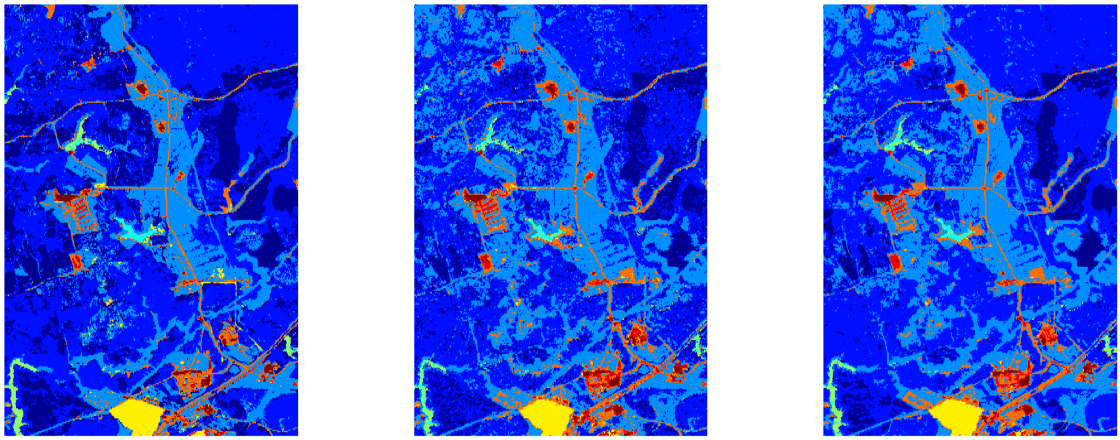


Fig. 2: Effect of the proposed coherence term on Indian Pines. From left to right: Ground-truth, classification with no spatial coherence, classification with spatial coherence, and reconstructed data (3×3, 20%) using a dictionary learned from the original data and spatial/spectral coherence. This is a color figure.



(a) (b) (c)

Fig. 3: Effect of coherence and significant sub-sampling in APHill mapping. (a) Original, (b) mapping after reconstructing 98% of the data ($3 \times 3, 2\%$) with no spatial coherence, and (c) mapping after reconstructing 98% of the data ($3 \times 3, 2\%$) with spatial/spectral coherence. This is a color figure.

the DMS method, which effectively uses spatial and spectral information from other pixels for a "grouping" effect in the abundance estimates, which is clearly shown in figures 2 and 3.

2) Classification accuracies are very similar to those obtained using the original datasets without sub-sampling, especially with the $3 \times 3, 20\%$, and $4 \times 4, 20\%$ realizations, thus showing that it is possible to get highly accurate classification performance with only a fifth of the data. It is worth mentioning that the decrease in classification accuracy is not due to gross errors in the detection ability of the method, but rather comes from errors involving highly similar classes, or insufficient training samples. For instance, focusing on datasets

reconstructed from only 2% of the data and a 3×3 spatial window, we observed that in the APHill scene, most of the misclassified pixels correspond to the "Lake1" and "Lake2" classes, with a 54.17% and 73.50% accuracy, respectively. From these classes, only 5 pixels were labeled from a class different than the "Lake1" or "Lake2" class. Similar errors occurred with the "Coniferous" and "Deciduous" classes, and the "Concrete" and "Road" classes. In the Urban scene, the two classes with the lowest performance for the DM method were "Road" and "Concrete", with 80% and 84.1121% respectively, with only 14 samples with misclassification errors from a different class.

The results shown in Table IV were obtained using training samples coming from the reconstructed images used for validation. In a more realistic approach, we provide classification results for cases in which the learning process was done *a priori* using data from the original dataset (no sub-sampling), and the validation is done on the reconstructed dataset. The idea is to show that a pre-learned dictionary (which is a compact representation of the classes) could be used in future acquisitions. These results are summarized in Table V for the HSI scenes after eliminating 80% of the data.

Finally, we have also tested completely eliminating 10% of the bands, simulating for example a sensor defect, where the optical arrays for several wavelengths are damaged. In addition to having missing bands, a large percentage of the rest of the data was removed at random as before. The results are summarized in Table VI. This shows that we can miss both entire bands and significant amounts of data at random, and still obtain very reasonable classification results, out-performing all the other tested methods.

D. Intermezzo

We make the following conclusions about the (supervised) experiments shown above:

1) The proposed method is superior in terms of classification performance when compared with standard classification approaches.
2) Accurate material classification can be obtained even when the majority of the data is missing. This is a clear example of how the data redundancy in high dimensional HSI can actually be beneficial for several tasks, including less expensive acquisition and faster transmission (at the cost of more sophisticated post-processing).

We now extend this to the unsupervised case, where there is no data for pre-training of the dictionary.

V. Unsupervised Mapping with a Block-Incoherent Dictionary

In previous sections we showed a methodology for mapping HSI when we know *a priori* the classes of interest (supervised mapping). The first stage consisted of learning per-class sub-dictionaries, and the second stage consisted of estimating the corresponding abundances. In this section we address the case for which there is no *a priori* information about the sources present in the scene, which could be seen as a blind source separation problem (unsupervised). A significant amount of research has been dedicated to automatically determining these endmembers. Many of the algorithms are simplex-projection based, where the vertices of the volume enclosing the data are considered as the endmembers [43]. The classical endmember estimation algorithms assume that there are pure

pixels in the image. Examples of this are the N-Finder [44], Vertex Component Analysis (VCA) [45], and Pixel Purity Index (PPI) [46]. Algorithms that do not make this assumption include Dependent Component Analysis (DECA) [47], Minimum Volume Simplex Analysis (MVSA) [48], and more recently, the Simplex Identification by Split Augmented Lagrangian (SISAL) method [49], which breaks the unmixing problem into smaller convex (and simpler) problems.

A second class of algorithms addressing this problem is based on nonnegative matrix factorization (NMF), attempting to find the matrices $\mathbf{\Psi}$ and $\mathbf{A}$ such that $\mathbf{Y} \approx \mathbf{\Psi A}$. Constrained versions of NMF for HSI unmixing have been extensively proposed, e.g., [50]–[52]. In addition, works that incorporate a sparsity constraint in NMF have been proposed in [53], [54], also in combination with a spatial constraint [55]. It is important to mention that in [55], the spatial constraint enforces similarities in the abundance vectors, while in our approach (DMS) we enforce local and non-local similarities that are weighted by spectral information. Minimum Volume Constrained (MVCNMF) [56], is a method for which the NMF is constrained to have minimum volume (thus limiting the non-uniqueness of the NMF because of the inward force). Finally, Minimum Dispersion NMF [57] was proposed as a method to deal with very flat spectra, where the constraint is to minimize the sum of the variances of the endmembers. In contrast with our proposed framework, these approaches force the materials to be represented by a single spectra, a very limiting model as explained before. In our method we let each data sample select the best possible sparse linear combination of atoms from a learned endmember sub-dictionary. As seen in the previous sections for supervised classification, representing a class/material by a single spectra is a limiting factor in terms of class reconstruction and accurate mapping. This motivates us to extend the above proposed dictionary based approach to the unsupervised case.

Our framework could be seen as a generalization of a sparsity and spatially constrained NMF, where the sources contain more than one atom per material. In order to automatically learn the sources online, just for the image being analyzed, and since the problems is non-convex, we need an initialization procedure such that the learned sub-dictionary for a given class maintains a relative separation from the sub-dictionaries learned for the other classes. We could naively set the number of atoms k_j per block and randomly select samples from the data, though doing this offers no guarantee of a valid sub-dictionary separation (and therefore separation between the sources). In this work, the sub-dictionaries are initialized by a fully constrained NMF (nonnegativity and sum to one of the abundance coefficients, CNMF). This sets the initial number of atoms per sub-dictionary to be $k_j = 1$ for all the classes j. Then, we will increase the number of atoms per sub-dictionary block as we progress in adapting the dictionary to the data. The idea is to eventually find the best class representation by increasing the dimensionality of the space where the class lives (represented by the sub-dictionary). It is intuitive that increasing this dimensionality will characterize the material more appropriately. Nevertheless, as the number of atoms per sub-dictionary increases, the cross-correlation (or coherence) of the dictionary, and in particular between the different sub-dictionaries, will increase. This has a negative impact in classification, especially with classes that are very close to each other. In addition, it has been shown in the sparse modeling literature that dictionary incoherence plays a crucial role in obtaining the correct sparse representation [20]. This is, the dictionary needs to be as incoherent as possible to get

a proper sparse representation of the signal, noting that if the coherence is maximum, any subset selection from the dictionary would yield similar solutions, hence losing uniqueness and discriminative power. To overcome this, we impose an additional constraint for keeping the different blocks of the dictionary as "pure" as possible by explicitly imposing a degree of incoherence. This is detailed next.

A. *Imposing Dictionary Incoherence*

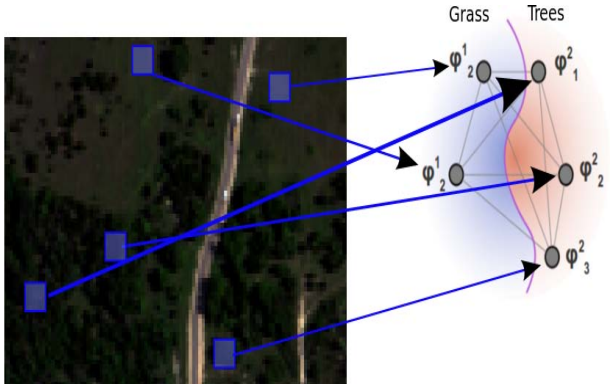


Fig. 4: Effect of coherence in the learned dictionary. The lower the coherence, the more separation between sub-dictionaries, and thus stronger discriminative power. This is a color figure.

The (sub-)dictionary incoherence plays the role of keeping the estimated sub-dictionary (block) for a given class as orthogonal as possible with each sub-dictionary corresponding to the other classes. As observed from the illustration in Figure 4, in order to perform classification, the sub-dictionaries representing the classes "Grass" and "Trees" need to maintain a certain degree of separation, since there are common features in the spectra that are shared among the classes (e.g., high amplitude in the visible/near-infrared regions). As the number of atoms per dictionary block increases, the more susceptible the dictionary becomes to these "mixing effects." To address this, we define a sub-dictionary incoherence term as [58]

$$\mathcal{I}(\mathbf{\Psi}^j) = \sum_{i \neq j}^{C} \|\mathbf{\Psi}^{iT}\mathbf{\Psi}^j\|_F^2. \tag{11}$$

With this we encourage the sub-dictionary $\mathbf{\Psi}^j$ to be separated from the rest of the sub-dictionaries. Since the atoms will not be completely orthogonal (e.g., due to the overcompletness of $\mathbf{\Psi}$), we penalize using (11). To update each of the dictionary atoms $\boldsymbol{\psi}_i^j$, we use the following update rule:

$$\begin{aligned} \boldsymbol{\psi}_i^{j,t} &\leftarrow P\{\boldsymbol{\psi}_i^{j,t-1}\} \\ &= \max(0, \boldsymbol{\psi}_i^{j,t-1} + \mu \frac{\partial}{\partial \boldsymbol{\psi}_i^j} \mathcal{R}_{\ell_2} + \mathcal{I}(\mathbf{\Psi})) \\ &= \max(0, \boldsymbol{\psi}_i^{j,t-1} + 2\mu((\mathbf{\Psi}^j\mathbf{A}^j - \mathbf{Y}^j)\mathbf{A}_i^{jT} + \lambda_I(\overline{\mathbf{\Psi}\mathbf{\Psi}}^T)\boldsymbol{\psi}_i^{j,t-1}), \end{aligned} \tag{12}$$

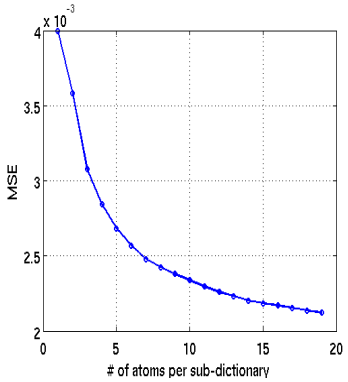


Fig. 5: Reconstruction Mean Square Error (MSE) as a function of the number of atoms per sub-dictionary for a single run in the Urban dataset.

where $\overline{\mathbf{\Psi}}$ is the concatenation of the sub-dictionary blocks not including the class being updated, $\mathbf{A}_i^j$ is the $i-th$ row of the coefficients matrix $\mathbf{A}^j$, and λ_I is the penalty parameter related to the sub-dictionaries incoherence. Again, each atom is normalized to have unit norm. Note that we update the atoms using the analyzed data $\mathbf{Y}^j$ (unsupervised case). In the supervised case, $\mathbf{Y}^j$ included the samples representing the $j-th$ class, and this was known *a priori*. Since we do not have training samples, we set during the iterations $\mathbf{Y}^j$ as the set of samples with maximum ℓ_1 norm in the corresponding class coefficients. For example, we initialize our algorithm using CNMF. Then, we split the data by assigning to $\mathbf{Y}^j$ the data samples whose largest abundance coefficients correspond to the $j-th$ class. We then proceed to update the dictionaries using this assignment following (12), and once these are updated, we re-assign the data following the same criteria. This is in the style of classical K-means, but with sub-dictionaries as "centroids," a different update rule (12), and an ℓ_1 criteria for assignment.

B. Stopping Criteria

In our scheme, we initialize the sparse modeling process using CNMF, and we progress the class representation by adding an atom (initialized at random) to each sub-dictionary. We desire the best possible class representation, so we stop adding atoms when the reconstruction error stops decreasing significantly,

$$(\{\frac{1}{n}\|\mathbf{Y}-\mathbf{\Psi}\mathbf{A}\|_F^2\}^{k_j}-\{\frac{1}{n}\|\mathbf{Y}-\mathbf{\Psi}\mathbf{A}\|_F^2\}^{k_j+1})\leq\eta, \tag{13}$$

where η is a threshold specified by the user. This decrease in MSE is illustrated in Figure 5, where after about 6 atoms the reconstruction error stops changing drastically. The algorithm for unsupervised mapping is summarized in Figure 6.

VI. Unsupervised Source Separation Experiments

In this section we test and compare our proposed algorithm with simulated and real HSI datacubes.

Input: Hyperspectral scene $\mathbf{Y}$, sparsity parameter λ_S, coherence parameter λ_G, sub-dictionary incoherence parameter λ_I, and stopping threshold η.

Output: Sparse matrix of fractional abundances $\mathbf{A}$ for $\mathbf{y}_i, \quad i = 1, ...n$, and learned endmember dictionary $\mathbf{\Psi}$.

Initialization

- Perform a Constrained Nonnegative Matrix Factorization (CNMF), $k_j = 1$ for all the classes j,

$$(\mathbf{A}^*, \mathbf{\Psi}^*) = \arg \min_{\mathbf{A} \succeq 0, \mathbf{1}^T\mathbf{A} = \mathbf{1}^T, \mathbf{\Psi} \succeq 0} \|\mathbf{Y} - \mathbf{\Psi}\mathbf{A}\|_F^2.$$

Simultaneous Learning

While (13) is not satisfied:

- Set $k_j \leftarrow k_j + 1$ (for all the classes) and initialize new atoms $\boldsymbol{\psi}^j$ randomly.
- Set Y^j as all data samples with largest abundance ℓ_1 norm corresponding to Ψ^j.
- Dictionary Update Stage:

$$\min_{\mathbf{\Psi}^j \succeq 0} \sum_{i=1}^{n_j} \|\mathbf{y}_i^j - \mathbf{\Psi}^j \boldsymbol{\alpha}_i^j\|_2^2 + \lambda_S \sum_{i=1}^{n_j} \mathcal{S}(\boldsymbol{\alpha}_i^j) + \lambda_G \sum_{i=1}^{n_j} \mathcal{G}(\boldsymbol{\alpha}_i^j, \mathbf{w}_i) + \lambda_I \mathcal{I}(\mathbf{\Psi}^j)$$

- Abundance Mapping Stage:

$$\min_{\boldsymbol{\alpha}_i \succeq 0} \sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{\Psi}\boldsymbol{\alpha}_i\|_2^2 + \lambda_S \sum_{i=1}^{n} \mathcal{S}(\boldsymbol{\alpha}_i) + \lambda_G \sum_{i=1}^{n} \mathcal{G}(\boldsymbol{\alpha}_i, \mathbf{w}_i).$$

Fig. 6: Algorithm for sub-pixel unsupervised classification in HSI.

A. Simulated HSI Data

We perform a series of experiments to compare the performance of the proposed algorithm with recently developed unmixing schemes:

1) A fully constrained NMF (CNMF, see references above). This is a standard fully constrained Nonnegative Matrix Factorization, where the sum to one constraint in the abundance coefficients is strictly enforced. This will also serve as our initialization algorithm as explained above.
2) Vertex Component Analysis (VCA) [45]. This is a minimum volume simplex approach. This algorithm assumes that there are pure pixels in the image. This serves as the initialization algorithm for SISAL below.
3) Simplex Identification via Split Augmented Lagrangian (SISAL) [49]. It does not assume pixel purity in the data.

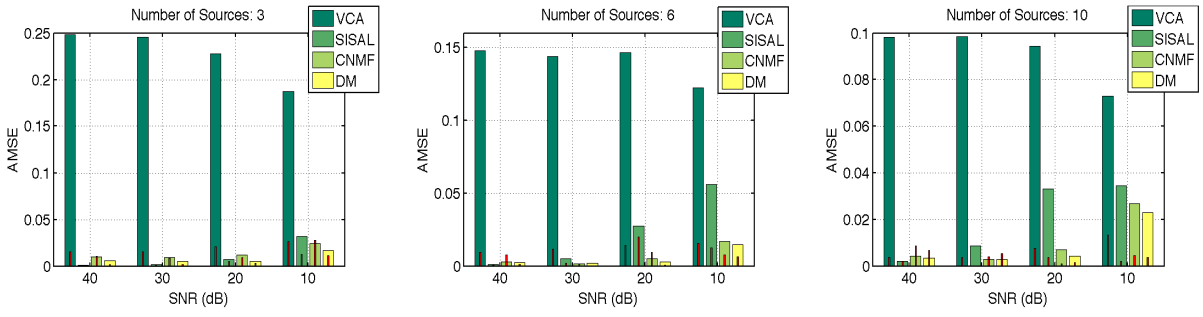
Both the VCA and SISAL algorithms come from the authors website. The code is publicly available at http://www.lx.it.pt/ bioucas/code.htm.

The simulated data is generated using the USGS spectral library available at http://speclab.cr.usgs.gov/spectral-lib.html, corresponding to the AVIRIS sensor, which consists of 500 mineral spectral signatures with 224 spectral bands. The "true" abundance values are generated following a Dirichlet probability density function (pdf) with density parameter of 0.1. This distribution guarantees nonnegativity and sum to one in the abundance coefficients. The dataset is generated with a total of $10,000$ observations, each pixel having a maximum purity of 0.8. The simulations test the algorithms using the following variants:

1) Additive noise was added to the dataset. The signal to noise ratio (SNR) is calculated as SNR $= 10 \cdot \log(\frac{\|\mathbf{X}\|_F^2}{\|\mathbf{N}\|_F^2})$, where $\mathbf{X}$ is the noiseless data. We include noise levels of 40, 30, 20, and 10 dB.
2) The number of sources was selected to be 3, 6, and 10.

Our measure of performance for this experiment is the mean squared error (MSE) between the original (ground truth) and computed abundance values, $\text{MSE}(\hat{\mathbf{A}}) = \frac{1}{n \cdot C}\|\mathbf{A} - \hat{\mathbf{A}}\|_F^2$. The parameters for the VCA and SISAL algorithms were chosen to be the default values in the public domain code released by the authors. For our algorithm we set the sparsity constraint as $\mathcal{S}(\boldsymbol{\alpha}) \leq 1$, and $\lambda_G = 0$. Each experiment was run 25 times, each time generating a datacube by drawing at random spectral signatures from the spectral library. The results are summarized in Figure 7. We have tested only the proposed DM, and not DMS, since for this simulated data there is no spatial information. We make the following observations:

1) While the algorithms perform relatively well, VCA performs poorly mainly due to the assumption about the presence of pure endmembers. In all cases, SISAL performs better than VCA.
2) The SISAL algorithm has the best performance under high SNR (40 dB). The proposed DM method performs better under lower SNRs and as the number of sources is increased. This is the case even when the algorithm is forced to have a fixed sparsity constraint. In addition, the data in this simulation assumes there is no spectral variability in the sources, while such variability will further favor our proposed technique. In all cases, DM performs better than CNMF.



(a) (b) (c)

Fig. 7: Tested algorithms performance in terms of abundance MSE for different values of SNR (dB). (a) 3 sources, (b) 6 sources, and (c) 10 sources. The green bars correspond to the average MSE, and the red bars to the error standard deviation. This is a color figure.

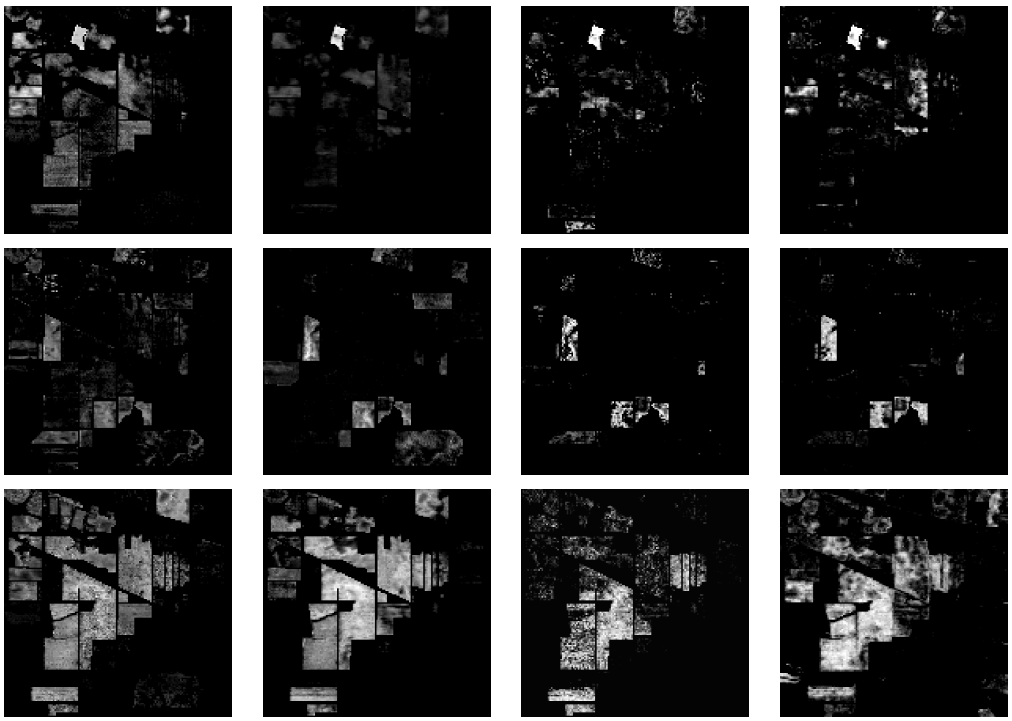
B. Experiments with Real HSI Data

We applied our proposed unsupervised algorithms DM and DMS to the three real HSI datacubes. In this experiment we show three abundance maps for each of the datacubes. We also included the results obtained by SISAL and CNMF (with parameters manually set to obtain the best results). For all images, we selected $\eta = 0.01$ and $\lambda_I = 0.01$ (parameters selected via cross-validation techniques). For the Indian Pines dataset, we show the results for the estimated "Stone-steel tower" class, the "Grass/Tress" class, and the "Soybeans-min" class. For the APHill dataset, we selected the estimated abundance values for the "Road" class, the "Coniferous" trees class, and the "Crop" class. Finally, the three abundance maps shown for the Urban data are the "Road," the "Soil," and the "Rooftop" class. While both CNMF and SISAL are very recent and powerful algorithms developed for HSI unmixing, still, our proposed approach showed advantages over these methods. Results are presented in figures (8), (9), and (10). These figures illustrate abundance maps obtained from the computed unmixing coefficients. Each pixel has a computed vector of coefficients (or abundances), and the presented mappings correspond to the values of the coefficients for a particular endmember. A dark pixel in the mapping means that the material is not present or has low presence, while a very bright pixel means that the material is very abundant. For SISAL and CNMF, the mappings correspond to the coefficient associated with each material (one spectra represents an endmember). For the proposed DM and DMS, the mapping corresponds to the ℓ_1 norm of the vector associated to the endmember sub-dictionary/class (ℓ_1 of the computed α^j).

We make the following observations from these experiments:

1) For the Indian Pines image, all the algorithms provided similar results in correctly identifying the stone tower, but incorrectly estimated high abundance values in other areas. Looking at the "Grass/Trees" class, the DMS algorithm was the best at identifying the locations of the class, and a clear advantage of the spatial coherence and a better fitting of the class by using dictionaries is observed. Finally, DM performed best at estimating the "Soybeans-min" class.
2) Looking at the APHill results, we observe that both DM and DMS gave the cleanest results. For example, the road estimation of SISAL include pixels pertaining to the lakes and some other wet areas. The CNMF wrongly estimated the "Crop" class region as "Road," an effect that is also seen in the "Crop" abundance estimates.
3) For the Urban image, we see mixing effects occurring with the "Road" estimation in both SISAL and CNMF algorithms, particularly with "Soil." Also, a high correlation in the abundance estimates of "Rooftop" and "Soil" is noticed with SISAL and NMF. Again, a clear advantage of a better class representation when using a learned structured dictionary instead of a single spectra is observed.

Given that we have training and validation data from each of the tested HSI datasets, we also conducted a numerical (quantitative) experiment to test the classification accuracy of our proposed unsupervised method. Since we know the locations of the "known" materials, we mapped the estimated sources from the unsupervised case with the corresponding classes from the supervised case. Table VII summarizes these results for APHill. There is a significant gain in classification accuracy when compared with the CNMF method (our initial condition), in



(a) (b) (c) (d)

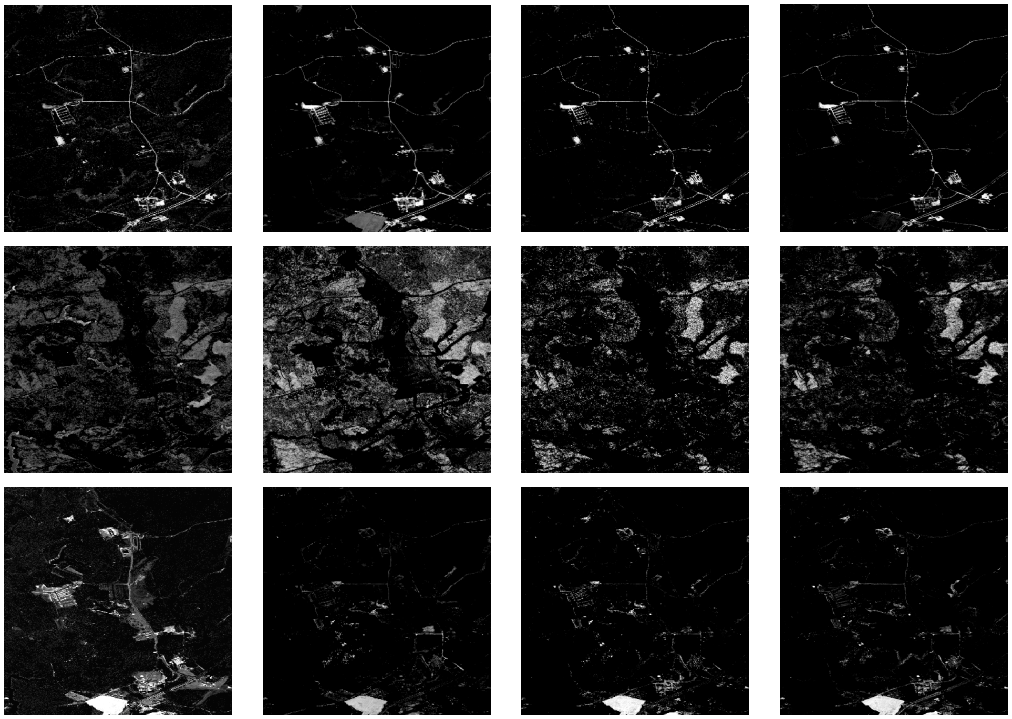
Fig. 8: Abundance maps corresponding to three classes from the Indian Pines dataset. The first row corresponds to the "Stone-steel tower" class, the second row corresponds to the "Grass/Trees" class, and the third row corresponds to the "Soybeans-min" class. (a) SISAL (b) NMF (c) DM (d) DMS.

particular for the "Road" class, for which most of the pixels pertaining to that class were confused with "Concrete" in the CNMF method. Similarly, most of the pixels corresponding to the "Grass" class are mislabeled as "Trees" using CNMF (see Figure 11). The only observed deterioration is for the confusion of some "Concrete" pixels with "Road" and "Gravel," three very similar classes. This is another example of the importance of a dictionary-based class representation and classification method, making the proposed DM/DMS a richer and more appropriate model to deal with high-dimensional HSI.

VII. Concluding Remarks

We have proposed supervised and unsupervised HSI composition mapping algorithms using learned sparse models. We reported the results on three hyperspectral datasets, and also showed the potential for a Bayesian compressed-sensing technique to help in solving acquisition, transmission, and storage issues related to HSI. With this framework, noise, class variability, and data redundancy are efficiently addressed by a structured sparse modeling classification technique that incorporates spatial/spectral regularization and class-dictionary incoherence.

While we currently increase the number of atoms uniformly for all sub-dictionaries in the unsupervised scenario, it will be interesting to do this class-dependent, letting different classes to learn different dictionary sizes. We plan to



(a) (b) (c) (d)

Fig. 9: Abundance maps corresponding to three classes from the APHill dataset. The first row corresponds to the "Road" class, the second row corresponds to the "Coniferous" class, and the third row corresponds to the "Crop" class. (a) SISAL (b) NMF (c) DM (d) DMS.

further exploit the proposed framework by considering adaptive sampling techniques, where sub-sampling is avoided in areas of limited spatial coverage or high spectral uncertainty. We are also further exploring the capabilities of the proposed framework to deal with the scenario where entire spectral bands or regions are missing, the preliminary results here reported are very encouraging. Lastly, we are investigating the possibility to classify directly from the sub-sampled data, without the need for pre-reconstruction, which is very important for real time applications [59].

Acknowledgments: Work partially supported by NGA, ONR, ARO, NSF, and AFOSR (NSSEFF). The authors would like to thank Dr. J. Duarte-Carvajalino, P. Sprechmann, Dr. O. Kuybeda, I. Ramirez, and F. Couzinie for very insightful and helpful discussions. We also thank J. Mairal and Dr. Bioucas-Dias for providing publicly available code used in this work.

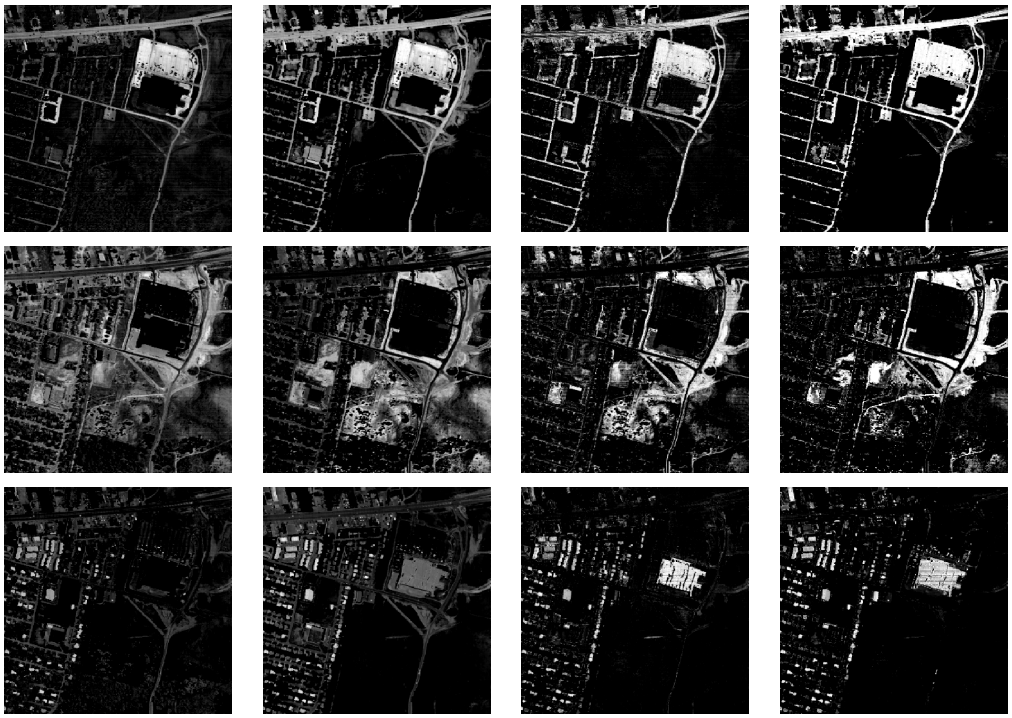
REFERENCES

[1] C.-I. Chang, *Hyperspectral Data Exploitation: Theory and Applications*. John Wiley and Sons, 2007.

[2] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. John Wiley and Sons, 2003.

[3] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, p. 110122, 2009.

[4] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin, "Non-parametric bayesian dictionary learning for sparse image representations," in *NIPS*, 2009.



(a) (b) (c) (d)

Fig. 10: Abundance maps corresponding to three classes from the Urban dataset. The first row corresponds to the "Road" class, the second row corresponds to the "Soil" class, and the third row corresponds to the "Rooftop" class. (a) SISAL (b) NMF (c) DM (d) DMS.

[5] L. C. Sanders, J. R. Schott, and R. V. Raqueno, "A vnir/swir atmospheric correction algorithm for hyperspectral imagery with adjacency effect," *Remote Sensing of Environment*, vol. 77, pp. 1–11, 2001.

[6] C. Bateson, G. Asner, and C. Wessman, "Endmember bundles: a new approach to incorporating endmember variability into spectral mixture analysis," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 38, no. 2, pp. 1083–1094, 2000.

[7] J. Chen, X. Jia, W. Yang, and B. Matsushita, "Generalization of subpixel analysis for hyperspectral data with flexibility in spectral similarity measures," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 7, pp. 2165–2171, 2009.

[8] D. Landgrebe, "Hyperspectral image data analysis," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 17–28, 2002.

[9] L. Jimenez and D. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 37, no. 6, pp. 2653–2667, 1999.

[10] D. L. Donoho, "High-dimensional data analysis: the curses and blessings of dimensionality," in *American Mathematical Society Conf. Math Challenges of the 21st Century*, 2000.

[11] L. du Plessis, S. Damelin, and M. Sears, "Reducing the dimensionality of hyperspectral data using diffusion maps," in *IGARSS*, vol. 4, 2009, pp. IV–885–IV–888.

[12] C. Bachmann and T. Ainsworth, "Bathymetric retrieval from manifold coordinate representations of hyperspectral imagery," in *IGARSS*, 2007, pp. 1548–1551.

[13] C. Bachmann, T. Ainsworth, and R. Fusina, "Improved manifold coordinate representations of large-scale hyperspectral scenes," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 10, pp. 2786–2803, 2006.

[14] Y. Zhou, B. Wu, D. Li, and R. Li, "Edge detection on hyperspectral imagery via manifold techniques," in *WHISPERS*, 2009, pp. 1–4.

[15] D. Tuia, G. Matasci, G. Camps-Valls, and M. Kanevski, "Learning the relevant image features with multiple kernels," in *IGARSS*, vol. 2, 2009, pp. II–65–II–68.

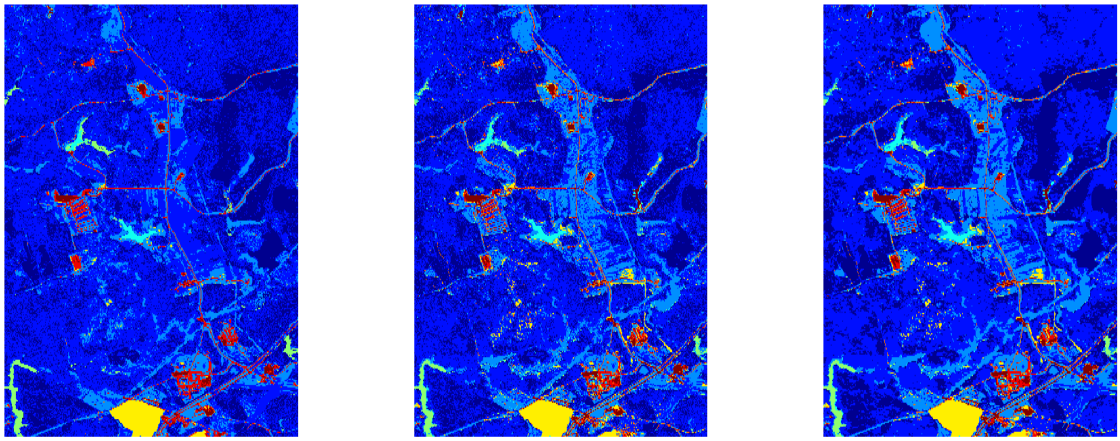


Fig. 11: Performance of the unsupervised algorithms in terms of classification accuracy. From left to right: CNMF, DM, and DMS. Each color represents an estimated material matching with the training and validation samples' coordinates (known *a priori*). See Table I and Figure 1 for the color code and class labels. This is a color figure.

[16] A. Mohan, G. Sapiro, and E. Bosch, "Spatially coherent nonlinear dimensionality reduction and segmentation of hyperspectral images," *Geoscience and Remote Sensing Letters, IEEE*, vol. 4, no. 2, pp. 206–210, 2007.

[17] Y. Bengio, J. F. Paiement, and P. Vincent, "Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering," in *Departement d'informatique et Recherche Operationnelle, Universite de Montreal*, 2003.

[18] D. Hsu, S. M. Kakade, J. Langford, and T. Zhang, "Multi-label prediction via compressed sensing," *http://arxiv.org/abs/0902.1284*, 2009.

[19] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *27th Asilomar Conf. on Signals, Systems and Comput.*, 1993.

[20] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[21] S. S. Chen, D. L. Donoho, Michael, and A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.

[22] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[23] P. Sprechmann and G. Sapiro, "Dictionary learning and sparse coding for unsupervised clustering," in *ICASSP*, 2010.

[24] M. S. Lewicki and B. A. Olshausen, "Inferring sparse, overcomplete image codes using an efficient coding framework," in *NIPS*, 1997.

[25] K. Engan, S. O. Aase, and J. H. Husøy, "Frame based signal compression using method of optimal directions (MOD)," in *ISCAS*, 1999, pp. 1–4.

[26] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[27] M. Elad and M. Aharon, "Image denoising via learned dictionaries and sparse representation," in *CVPR*, 2006, pp. 17–22.

[28] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *NIPS*, 2007, pp. 801–808.

[29] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *CVPR*, 2008, pp. 1–8.

[30] ——, "Supervised dictionary learning," in *NIPS*, 2008, pp. 1033–1040.

[31] J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce, "Discriminative sparse image models for class-specific edge detection and image interpretation," in *ECCV*, 2008.

[32] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *ICML*, 2007.

[33] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *ICCV*, 2009.

[34] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2008.

[35] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[36] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[37] N. Keshava and J. Mustard, "Spectral unmixing," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 44–57, 2002.

[38] Z. Guo and S. Osher, "Template matching via l1 minimization and its application to hyperspectral target detection," UCLA, www.math.ucla.edu/applied/cam/, Tech. Rep. 09-103, 2009.

[39] Z. Guo, T. Wittman, and S. Osher, "L1 unmixing and its applications to hyperspectral image enhancement," UCLA, www.math.ucla.edu/applied/cam/, Tech. Rep. 09-30, 2009.

[40] M. Velez-Reyes and S. Rosario, "Solving adundance estimation in hyperspectral unmixing as a least distance problem," in *IGARSS*, vol. 5, 2004, pp. 3276–3278.

[41] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *CVPR*, vol. 2, 2005, pp. 60–65.

[42] G. Camps-Valls, S. Member, B. Tatyana V., and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, pp. 2044–3054, 2007.

[43] M. Craig, "Minimum-volume transforms for remotely sensed data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 32, no. 3, pp. 542–552, 1994.

[44] E. M. Winter and M. E. Winter, "Autonomous hyperspectral end-member determination methods," *Sensors, Systems, and Next-Generation Satellites III*, vol. 3870, no. 1, pp. 150–158, 1999.

[45] J. Nascimento and J. Dias, "Vertex component analysis: a fast algorithm to unmix hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 4, pp. 898–910, 2005.

[46] J. Boardman, F. Kruse, and R. Green, "Mapping target signatures via partial unmixing of aviris data," in *Summaries of JPL Air-borne Earth Science Workshop*, 1995.

[47] J. M. P. Nascimento and J. M. Bioucas-Dias, "Dependent component analysis: A hyperspectral unmixing algorithm," in *IbPRIA (2)*, 2007, pp. 612–619.

[48] J. Li and J. Bioucas-Dias, "Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data," in *IGARSS*, vol. 3, jul. 2008, pp. III–250–III–253.

[49] J. Bioucas-Dias, "A variable splitting augmented Lagrangian approach to linear spectral unmixing," *http://arxiv.org/abs/0904.4635*, 2009.

[50] Y. Masalmah and M. Velez-Reyes, "Unsupervised unmixing of hyperspectral imagery," in *MWSCAS*, vol. 2, 2006, pp. 337–341.

[51] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," in *Computational Statistics and Data Analysis*, 2006, pp. 155–173.

[52] D. Heinz and Chein-I-Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, no. 3, pp. 529–545, 2001.

[53] P. O. Hoyer and P. Dayan, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[54] A. Zare and P. Gader, "Sparsity promoting iterated constrained endmember detection with integrated band selection," in *IGARSS*, jul. 2007, pp. 4045–4048.

[55] S. Jia and Y. Qian, "Spectral and spatial complexity-based hyperspectral unmixing," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 12, pp. 3867–3879, 2007.

[56] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 3, pp. 765–777, 2007.

[57] A. Huck, M. Guillaume, and J. Blanc-Talon, "Minimum dispersion constrained nonnegative matrix factorization to unmix hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 6, pp. 2590–2602, 2010.

[58] I. Ramirez, F. Lecumberry, and G. Sapiro, "Universal priors for sparse modeling," in *The Third International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, Aruba*, 2009.

[59] Q. Du and R. Nekovei, "Fast real-time onboard processing of hyperspectral imagery for detection and classification," *J. Real-Time Image Processing*, vol. 4, pp. 273–286, 2009.

TABLE I: Class labels and number of samples per class for the Indian Pines, APHill, and Urban HSI cubes.

Class		Samples	
Label	**Name**	**Train**	**Test**
Indian Pines			
1	Alfalfa	27	27
2	Cornnotill	717	717
3	Corn-min	417	417
4	Corn	117	117
5	Grass/Pasture	248	249
6	Grass/Trees	373	374
7	Grass/pasture-mowed	13	13
8	Hay-windrowed	244	245
9	Oats	10	10
10	Soybeans-notill	484	484
11	Soybeans-min	1234	1234
12	Soybean-clean	307	307
13	Wheat	106	106
14	Woods	647	647
15	Bldg-Grass-Tree-Drives	190	190
16	Stone-steel towers	47	48
APHill			
1	Coniferous	597	598
2	Deciduous	1290	1290
3	Grass	829	829
4	Lake1	120	120
5	Lake2	117	117
6	Crop	792	792
7	Road	123	123
8	Concrete	49	50
9	Gravel	62	63
Urban			
1	Road	254	255
2	Concrete	106	107
3	Dark soil	76	76
4	Bright soil	39	40
5	Gray rooftop	417	417
6	Brown rooftop	95	96
7	Grass	338	339
8	Trees	276	277

TABLE II: Results for the first supervised classification experiment. Shown are the mean (first row) and standard deviation (second row) of 25 runs for three HSI datasets. Best results are in bold.

Image	**Accuracy (mean and standard deviation)**				
	ED	**SAM**	ℓ_1	**DM**	**DMS**
Indian Pines	0.3788	0.4437	0.8639	0.878	**0.9352**
	0.0082	0.006	0.0045	0.0048	0.0038
APHill	0.9006	0.9151	0.9758	0.9894	**0.9966**
	0.0034	0.0031	0.0023	0.0011	0.0009
Urban	0.9013	0.972	0.9894	0.999	**0.9992**
	0.0047	0.0034	0.0032	0.0009	0.0007

TABLE III: PSNR and spectral angles between the original and reconstructed datasets. The first column shows the parameters selected for the reconstruction, the spatial window size and the amount of available data used for reconstruction. The spectral angles are in degrees.

Image	**Reconstruction PSNR and Spectral Angle**				
Indian Pines	**PSNR**	**min**	**max**	**avg.**	**median**
$3 \times 3, 2\%$	35.1827	0.6558	27.9661	2.6519	2.2808
$3 \times 3, 5\%$	42.0383	0.3353	21.9318	1.2400	1.0551
$3 \times 3, 10\%$	48.0072	0.2864	8.1674	0.7328	0.6849
$3 \times 3, 20\%$	50.0940	0.2694	8.2588	0.6186	0.5964
$4 \times 4, 2\%$	36.5539	0.5219	25.9245	2.1256	1.7854
$4 \times 4, 5\%$	42.7197	0.3121	18.8338	1.1327	0.9553
$4 \times 4, 20\%$	50.0670	0.2856	8.3672	0.6197	0.5952
$5 \times 5, 2\%$	36.4205	0.4386	27.2215	2.0617	1.6809
APHill					
$3 \times 3, 2\%$	33.2542	0.5492	56.9784	2.7071	2.0083
$3 \times 3, 5\%$	38.2726	0.2923	74.2222	1.2897	0.9920
$3 \times 3, 10\%$	43.8655	0.2559	22.6844	0.9632	0.7928
$3 \times 3, 20\%$	46.1789	0.3563	14.1513	1.1429	0.9838
$4 \times 4, 2\%$	37.8549	0.4632	74.1300	2.4906	1.8053
$4 \times 4, 5\%$	40.0584	0.3383	65.6407	1.6301	1.2753
$4 \times 4, 20\%$	46.2491	0.3196	19.6792	1.3074	1.0892
$5 \times 5, 2\%$	33.5847	0.4393	72.7613	2.4333	1.7413
Urban					
$3 \times 3, 2\%$	30.3053	2.5799	61.4310	7.8448	7.1313
$3 \times 3, 5\%$	38.8684	2.4565	32.9800	6.3067	6.2368
$3 \times 3, 10\%$	43.0309	2.4738	26.7932	6.0401	6.0933
$3 \times 3, 20\%$	46.2861	0.4358	21.9183	1.5765	1.2608
$4 \times 4, 2\%$	32.4952	0.6723	59.0952	4.8124	3.7554
$4 \times 4, 5\%$	40.6783	0.6273	37.3109	2.4521	1.9213
$4 \times 4, 20\%$	45.6692	0.4387	23.1958	1.6735	1.3312
$5 \times 5, 2\%$	32.8143	0.7574	59.6115	4.5828	3.5597

TABLE IV: Overall classification accuracies for the datasets reconstructed from highly under-sampled data. The first column shows the spatial window and data percentage used for reconstruction.

Image	**Accuracy**				
	ED	**SAM**	ℓ_1	**DM**	**DMS**
Indian Pines					
$3 \times 3, 2\%$	0.3823	0.3086	0.5070	0.4426	**0.7666**
$3 \times 3, 5\%$	0.3742	0.4177	0.7477	0.7381	**0.8781**
$3 \times 3, 10\%$	0.3797	0.4332	0.8216	0.8141	**0.9001**
$3 \times 3, 20\%$	0.3796	0.4359	0.8339	0.8253	**0.9007**
$4 \times 4, 2\%$	0.3819	0.3767	0.6901	0.6650	**0.8785**
$4 \times 4, 5\%$	0.3821	0.4272	0.8006	0.8021	**0.9014**
$4 \times 4, 20\%$	0.3807	0.4380	0.8467	0.8287	**0.9115**
$5 \times 5, 2\%$	0.3801	0.3813	0.7477	0.7724	**0.8986**
APHill					
$3 \times 3, 2\%$	0.8920	0.8544	0.8976	0.9111	**0.9822**
$3 \times 3, 5\%$	0.8885	0.9121	0.9681	0.9736	**0.9917**
$3 \times 3, 10\%$	0.8933	0.9119	0.9751	0.9772	**0.9917**
$3 \times 3, 20\%$	0.8943	0.9149	0.9782	0.9792	**0.9915**
$4 \times 4, 2\%$	0.8991	0.8890	0.9455	0.9546	**0.9897**
$4 \times 4, 5\%$	0.8933	0.9169	0.9819	0.9804	**0.9950**
$4 \times 4, 20\%$	0.8943	0.9232	0.9862	0.9874	**0.9932**
$5 \times 5, 2\%$	0.9028	0.9064	0.9628	0.9739	**0.9930**
Urban					
$3 \times 3, 2\%$	0.9085	0.8836	0.9334	0.9496	**0.9944**
$3 \times 3, 5\%$	0.8911	0.9452	0.9907	0.9963	**0.9981**
$3 \times 3, 10\%$	0.8930	0.9645	0.9919	0.9969	**0.9994**
$3 \times 3, 20\%$	0.9042	0.9701	0.9894	0.9981	**1.0000**
$4 \times 4, 2\%$	0.9123	0.9011	0.9577	0.9807	**0.9969**
$4 \times 4, 5\%$	0.9048	0.9564	0.9944	0.9988	**0.9994**
$4 \times 4, 20\%$	0.9048	0.9708	0.9888	0.9988	**1.0000**
$5 \times 5, 2\%$	0.9135	0.8955	0.9664	0.9907	**0.9975**

TABLE V: Overall classification accuracy for reconstructed data with only 20% of the pixels used for reconstruction. The class dictionaries were learned *a priori* using the original dataset.

Image	**Accuracy**				
	ED	**SAM**	ℓ_1	**DM**	**DMS**
Indian Pines	0.3720	0.4432	0.7811	0.7967	**0.8434**
APHill	0.8353	0.9269	0.9397	0.9538	**0.9691**
Urban	0.8936	0.8955	0.9589	0.9670	**0.9844**

TABLE VI: Classification results for the Urban dataset when entire bands are missing in addition to the missing data at random as before. The data is reconstructed from this highly under-sampled data before classification. The first column shows the spatial window, data percentage, and percentage of bands used for reconstruction.

Image	**Accuracy**				
	ED	**SAM**	ℓ_1	**DM**	**DMS**
Original	0.9079	0.9850	0.9850	0.9981	**1.0000**
$2 \times 2, 2\%, 90\%$	0.8609	0.8571	0.8797	0.9117	**0.9568**
$2 \times 2, 5\%, 90\%$	0.8966	0.9586	0.9718	0.9925	**0.9962**
$4 \times 4, 2\%, 90\%$	0.8910	0.8835	0.9305	0.9718	**0.9699**
$4 \times 4, 5\%, 90\%$	0.8947	0.9380	0.9474	0.9774	**0.9906**

TABLE VII: Unsupervised per-class overall classification accuracy for the APHill dataset based on the training and validation data used in the supervised case. See Table I for the class labels.

Method	**Class label**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
CNMF	92.5523	63.6047	17.0688	**100.0000**	**100.0000**	**100.0000**	1.2146	**97.9798**	92.0000
DM	92.8870	78.9535	67.1894	**100.0000**	**100.0000**	**100.0000**	55.0607	62.6263	98.4000
DMS	**99.7490**	**91.7054**	**73.8239**	**100.0000**	**100.0000**	**100.0000**	**63.5628**	76.7677	**99.2000**